164-TP-001-002

Methodology for Estimating DAAC-to-DAAC Traffic Due to User Pull

Technical Paper--Not intended for formal review or government approval.

May 1996

Prepared Under Contract NAS 5-60000

RESPONSIBLE SCIENTIST/ENGINEER

Lori J. Tyahla /s/ 05/03/96

Lori J. Tyahla, Science Specialist II, ESSi Date
EOSDIS Core System Project

SUBMITTED BY

Joy Colucci /s/ 05/15/96

Dr. Joy Colucci, Manager, Science Office Date
EOSDIS Core System Project

Hughes Information Technology Systems
Upper Marlboro, Maryland

This page intentionally left blank.

Abstract

A basic methodology is presented for providing a rough estimate for the volume of the Inter-DAAC traffic resulting from user queries to EOSDIS. This methodology was originally developed to aid in estimation of the required network bandwidth between DAACs. The ECS User Characterization Science User Scenarios, User Pull Technical Baseline, and the EOSDIS Product Use Survey were the resources utilized in this methodology.

keywords: network traffic, DAAC, user pull, scenarios, baseline, survey

This page intentionally left blank.

1. Introduction

1.1 Purpose

The purpose of this document is to describe the methods used by the ECS User Characterization Team to estimate the volume of DAAC-to-DAAC traffic resulting from user queries. This data is provided to ECS developers to support the design of the DAAC-to-DAAC communications network.

1.2 Organization

This document consists of two main sections. Section 2 details the sources of input data used in the analysis. Section 3 presents the analysis methods and results.

1.3 Review and Approval

This Technical Paper is an informal document approved at the ECS Office Manager level. It does not require formal Government review or approval; however, it is submitted with the intent that review and comments will be forthcoming.

The ideas expressed in this Technical Paper are valid for six months from the approval date. Questions concerning distribution or control of this document should be addressed to:

Data Management Office
The ECS Project Office
Hughes Information Technology Sytems
1616 McCormick Drive
Upper Marlboro, MD 20774

2. Inputs and Assumptions

In order to determine the required network bandwidth between DAACs, it is necessary to estimate the rate at which data will be flowing into and out of each DAAC. There are several contributors to this traffic; one major contributor is the exchange of data between DAACs for the purpose of producing a higher level product at the destination DAAC. It has generally been assumed that the volume of inter-DAAC traffic due to user queries is much smaller than that due to data processing, but the user query volume is still estimated and accounted for in the overall traffic total. The methodology described here provides a rough estimate for the volume of inter-DAAC traffic resulting only from user queries to the EOSDIS.

2.1 Inputs and Assumptions

The inputs to the methodology described in this analysis are the science user scenarios collected from 27 science users, the User Pull Technical Baseline (2/5/96 #) and the results of the EOSDIS Product Use Survey (# 161-TP-001-001). The timeframes under consideration are: early 1998, early 1999, mid-1999, and Jan. 1, 2000.

2.1.1 Science User Scenarios

The ECS User Characterization Team (UCT) developed a method for categorizing the user community according to system access patterns and geographic scale of research (for details, see *ECS User Characterization Methodology and Results*, Sept., 1994, #194-00313TPW). This categorization is represented as a matrix with these parameters as the two principle components; the matrix is referred to as the Science User Scenario Matrix. The ECS User Characterization Team then interviewed 27 scientists and collected a "user scenario" (or "use case") from each describing how they would interact with the EOSDIS. Each scenario is a detailed, step-by-step description of the services each scientist would invoke and the data he or she would access in each step.

Each step in each scenario is then translated by the UCT to a system perspective. For example, when a user places an order for data (from the user's perspective, this is a data order), the request is decomposed into component subservices (the system must locate the data in the archive, retrieve it from the archive, subset it if necessary, and distribute it to the user). Thus, placing an order causes the following series of subservices to be invoked: "locate, retrieve, spatial subset, distribute". Currently, there are 65 of these mid-level subservices.

A number of "service invocations" occur when a user makes any type of request of the system; the results of the request may or may not be interactive (for example, when the user sets up a standing order for data, he is invoking a system subservice, "place order", but the result of this request is that data is delivered to the user automatically at some later date). In addition, some of the scenarios are enacted by the same user several times per year. Since the technical baseline information applies to a yearly timescale, the number of times per year that a scenario is enacted is a multiplicative factor in the number of times one particular user invokes services on a yearly basis.

2.1.2 EOSDIS Product Use Survey

The science scenarios provide information regarding the types of services that users will invoke and the rate at which the invocations will occur. However, the data products accessed by the scenarios do not span the entire list of available products. Thus, in 1995, the UCT designed a product survey and implemented it on the World Wide Web. The main purpose of the survey was to gauge interest in the data products that the EOSDIS will produce and archive. E-mail messages were sent to approximately 4,000 science users inviting them to complete the survey; other users discovered the survey on their own and completed it. The UCT received about 400 complete responses.

The survey responses were used to determine a Relative Product Access Frequency (RPAF) for each individual data product relative to the rest. Usually, the RPAFS are aggregated into Relative DAAC Access Frequencies (RDAFs) based upon the archive location of each product. These RDAFs are then used to determine the proportion of user requests arriving at each individual DAAC. For example, if the RDAF for Goddard Space Flight Center (GSFC) is 0.60, then 60% of all estimated user requests will be for products and services at GSFC.

2.1.3 User Pull Technical Baseline

The total number of science users was obtained from the ECS User Pull Technical Baseline. The numbers in the baseline were arrived at by using current data system usage statistics at the DAACs to project future usage. The result of the analysis provides the total number of expected science users in each of the four epochs listed in Section 1.1. For more information on the demographic analysis, see *ECS User Characterization Methodology and Results*, Sept., 1994, (#194-00313TPW).

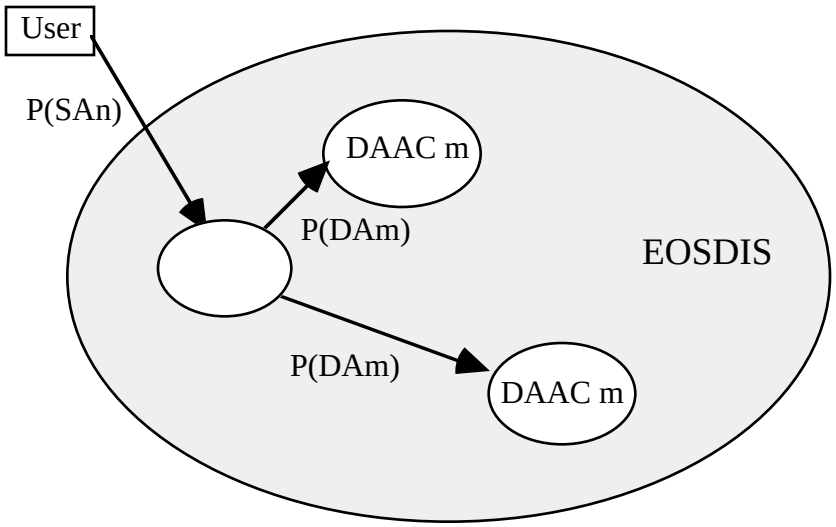
The number of ***system*** accesses through each DAAC per year is also presented in the User Pull Technical Baseline. The User Pull Technical Baseline reports the number of distinct users currently using each DAAC as well as the projected number of distinct users of each DAAC. This information is based on usage statistics at existing data centers. It is assumed that these users will access (or connect to) EOSDIS at the DAAC with which they are associated. An additional assumption is made regarding the proportion of users who are likely to invoke services against data at more than one DAAC to avoid double-counting users. The number of ***system*** accesses describes the ***number of connections*** to the EOSDIS - it does not describe the services and data that the users are accessing while they are connected. This process is explained in detail in Section 3.0.

3. Methodology

One can view the system in a generic way as represented below (not all DAACs as shown). A user connects to EOSDIS through a DAAC (a ***system*** access) and may or may not access other DAACs from the original DAAC. We can use the following terms:

$P(SA_n)$ = the probability that a user connects to EOSDIS via DAAC *n*

$P(DA_m)$ = the probability that a user accesses DAAC *m* from the DAAC that he or she is connected to (DAAC *n*)



$P(SA_n)$ is the probability that a user will connect to EOSDIS via DAAC *n*.. The values are based upon the User Pull Technical Baseline (2/5/96) as explained above. $P(DA_m)$ is the probability that a user will access DAAC *m* from DAAC *n*. These values are based upon the results of the EOSDIS Product Use Survey as described in Section 2.1.2. Then, the probability that a user will submit a query from GSFC to LaRC is:

$$P(A_{GSFC \rightarrow LaRC}) = P(SA_{GSFC}) \times P(DA_{LaRC})$$

Note that the volume of the query itself appears in the traffic flow from GSFC to LaRC and the volume of the query ***results*** contributes to the traffic flow from LaRC to GSFC.

It is assumed that all Browse and Level data (Levels 1 through 4) travel directly from the DAAC queried to the user (based on discussions with developers) and that the Guide is replicated at each DAAC. Also, volumes resulting from users connecting to EOSDIS through a DAAC and accessing data and services at that DAAC only are not included since they do not cause traffic between two DAACs.

3.1 Detailed Procedure

1. The scenario database was examined and, for each step where data or search results are delivered electronically to a user connected to the system through one of the DAACs , the volume of the query and the volume of the results are recorded. Also recorded is the number of times per year that the query is submitted and the number of users who submit similar queries.

2. From step 1, we have two volumes in MB yr^{-1} $user^{-1}$ per scenario step (the volume of the queries and the volume of the results) and a number of users for each scenario step of interest. Next, these volumes are multiplied by the number of users for each step of interest resulting in volumes yr^{-1} $step^{-1}$.

3. The results of step 2 are then summed over all scenario steps resulting in two total volumes of data (queries and results) that are moved across the network in one year (MB yr^{-1}).

4. Two separate tables were then constructed - one for query volumes and one for results volumes. The volume of traffic flowing between each possible two-DAAC combination (LaRC to GSFC is a different combination than GSFC to LaRC) due to the user queries was determined by multiplying the total yearly volume of the user queries by the access probabilities for each DAAC combination. For example, the volume of traffic from GSFC to LaRC is determined by:

$$\text{Query Vol (GSFC} \rightarrow \text{LaRC)} = \beta\, P(A_{GSFC \rightarrow LaRC}) = \beta\, P(SA_{GSFC}) \times P(DA_{LaRC})$$

   where β is the total yearly volume of all user queries. The volume due to results of user queries is determined in the same way, taking care to preserve directional information. For example, the volume of results due to user queries from GSFC to LaRC actually flows from LaRC to GSFC:

$$\text{Results Vol (LaRC} \rightarrow \text{GSFC)} = \lambda\, P(A_{GSFC \rightarrow LaRC}) = \lambda\, P(SA_{GSFC}) \times P(DA_{LaRC})$$

   where λ is the total volume of query results.

5. The final step is to sum the query volumes and the results volumes, again taking care to preserve the directional differences. For example, the results volume flowing from GSFC to LaRC is added to query volumes that flow from GSFC to LaRC.

4. Results

The results of the procedure described in this paper are shown below for 4 epochs. **All volumes are MB/year.**

Early 98						
Total Query (Query and Results) Volume going between DAACs						
	To: DAAC					
From: DAAC	ASF	EDC	GSFC	JPL	LaRC	NSIDC
ASF	0	1331.96	1075.159	1235.06	676.534	266.5078
EDC	576.253	0	1670.146	1849.269	1036.016	400.036
GSFC	4522.63	15814.5	0	14680.05	7936.274	3163.218
JPL	284.844	979.5613	862.6691	0	524.0815	196.2887
LaRC	1788.98	6252.058	4962.416	5802.403	0	1250.614
NSIDC	131.72	457.2723	379.3191	423.3781	236.024	0

Early 99						
Total Query (Query and Results) Volume going between DAACs						
	To: DAAC					
From: DAAC	ASF	EDC	GSFC	JPL	LaRC	NSIDC
ASF	0	1790.516	1445.31	1660.256	909.4467	358.2589
EDC	774.641	0	2245.142	2485.92	1392.691	537.7574
GSFC	6079.65	21258.98	0	19733.98	10668.51	4252.225
JPL	382.909	1316.798	1159.672	0	704.5121	263.8655
LaRC	2404.87	8404.464	6670.84	7800.005	0	1681.165
NSIDC	177.067	614.6984	509.9101	569.1353	317.2813	0

Mid 99						
Total Query (Query and Results) Volume going between DAACs						
	To: DAAC					
From: DAAC	ASF	EDC	GSFC	JPL	LaRC	NSIDC
ASF	0	1814.67	1464.817	1682.653	921.7188	363.0919
EDC	785.094	0	2275.465	2519.456	1411.493	545.0125
GSFC	6161.66	21545.76	0	20000.18	10812.43	4309.586
JPL	388.077	1334.565	1175.35	0	714.0292	267.4259
LaRC	2437.31	8517.838	6760.844	7905.224	0	1703.844
NSIDC	179.456	622.9913	516.796	576.8131	321.5642	0

Jan. 2000						
Total Query (Query and Results) Volume going between DAACs						
	To: DAAC					
From: DAAC	ASF	EDC	GSFC	JPL	LaRC	NSIDC
ASF	0	2003.228	1617.003	1857.494	1017.485	400.8197
EDC	866.666	0	2511.832	2781.245	1558.13	601.6419
GSFC	6801.91	23784.54	0	22078.37	11935.92	4757.387
JPL	428.396	1473.229	1297.411	0	788.1961	295.2119
LaRC	2690.57	9402.911	7463.318	8726.643	0	1880.887
NSIDC	198.102	687.7236	570.4809	636.7481	354.9717	0